

Fermi National Accelerator Laboratory

FERMILAB-Conf-91/351

Modeling of the DZero Data Acquisition System

R. Angstadt, M. Johnson and I. Manning

*Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510*

J. Wightman

*Dept. of Physics, Texas A&M University
College Station, Texas 77843*

December 1991

Presented at the *IEEE Nuclear Science Symposium*, Santa Fe, New Mexico, November 2 - 9, 1991.

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Modeling of the DZero Data Acquisition System

R. Angstadt, M. Johnson and I. L. Manning
Fermilab, # PO Box 500, Batavia, IL 60510, USA

J. A. Wightman

Dept of Physics, Texas A&M Univ, College Station, TX 77843
Texas Accelerator Center, The Woodlands, TX 77381

ABSTRACT

A queuing theory model was used in the initial design of the D0 data acquisition system. It was mainly used for the front end electronic systems. Since then the model has been extended to include the entire data path for the tracking system. The tracking system generates the most data so we expect this system to determine the overall transfer rate. The model was developed using both analytical and simulation methods for solving a series of single server queues. We describe the model and the methods used to develop it. We also present results from the original models, updated calculations representing the system as built and comparisons with measurements made with the hardware in place for the cosmic ray test run.

I. INTRODUCTION

The need for a model of the DZero data acquisition system became apparent very early in the construction of the detector. The hardware designers faced many choices such as the speed of the crate backplanes and the number of front end buffers. Many of these choices would be difficult or impossible to change later on if they were discovered to be a bottleneck in the data acquisition system.

We use queuing theory terminology in this paper. We define these terms with examples from everyday life. Thus a single server queue is a car wash with one wash bay. The queue size corresponds to the number of parking spaces in front of the car wash. A double buffered system is then a car wash with one parking space in front. The queue service time is the time to wash one car. In queuing theory these service times usually have a Poisson time distribution.

The data for this paper was taken with the hardware that was in place for the recently completed cosmic ray test run. The structure of the data acquisition system is similar to the final configuration so the models remain valid. However, substantially faster hardware is now being installed for the first physics run. We present modeling results for both systems.

Section II describes the DZero data acquisition system. Section III describes the analytical queuing theory models of the system. Results of the Monte Carlo model for different configurations and comparison with the present hardware system are described in section IV.

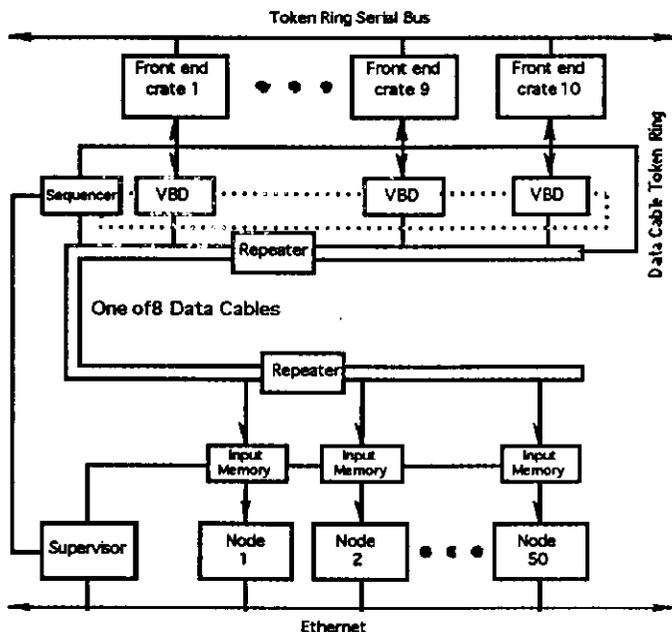


Figure 1. Block diagram of the data cable for the Vertex detector. The token is sent over the data cable; the dotted line indicates the token control bus. Each MicroVAX has dualport input memories - one for each cable.

II. DESCRIPTION OF THE DZERO DATA ACQUISITION SYSTEM

The DZero data acquisition system consists of 79 front end digitizer crates distributed over 8 parallel data highways or data cables. Figure 1 shows a block diagram of one of these data cables. Each of these crates has a serial connection to an IBM Token Ring Local Area Network and a parallel connection to the data cables[1]. The serial link is used for down loading and diagnostics while the parallel link is used for transferring data to the level 2(L2) trigger system which consists of 50 MicroVAX computers. Events are only sent to L2 if they pass a hardware trigger(L1)[2]. Each data cable is configured as a small token ring system (no relation to the IBM one). The crate controller linking a crate to one of the data cables is a VME Buffer/Driver (VBD). This device accepts a token from a readout controller (SEQUENCER). If it is ready to read out the data, it takes the token, transfers the data on to the cable and then releases the token. Otherwise the token is

immediately passed onto the next crate. The token continues to circulate until all crates have read out. This system reads the crates in the order that the data is ready for transfer, not in any fixed numerical order.

At the other end of the data cable are 50 MicroVAX computers. Each data cable is connected to a dual port memory on each of the MicroVAXs. A device called the trigger supervisor determines which MicroVAX should receive a given event and enables its set of 8 dual port memories. The data is then strobed into the dual port memory as it flows by on the data cable.

III. ANALYTICAL MODEL OF THE DATA ACQUISITION SYSTEM

A. Single Front End Crate

The modeling project was started with an analytical model for the flash ADC(FADC) system for the central tracking electronics. This system was chosen because it generates more than 75% of the DZero data. The first model consisted of a single crate with between 1 and 3 buffers and one VBD also with between 1 and 3 buffers. These correspond to queue lengths of 0,1 or 2. The service time is the transfer time across the backplane for the first queue and the transfer time out of the crate for the second queue. We have ignored such things as the time to digitize and zero suppress the data. The time between a trigger and data ready to transfer across the backplane is typically 15 microseconds while a typical backplane transfer time is 400 microseconds which is a ratio of 25. Similar ratios hold for the transfer time out of the VBD to the level 2 trigger.

The event arrival rate is assumed to be Poisson. The queue service times are determined by the event size since the backplane and VBD transfer speeds are fixed. Results from the DZero Monte Carlo program show that the event size distribution is Gaussian. We have used the Poisson distribution for our analytical calculations but the Poisson distribution approaches a Gaussian for large numbers so we expect little difference. Computer calculations using a Gaussian distribution support this.

Table 1 gives this distribution for the Vertex Chamber (VTX) which is believed to be the most heavily used data cable. It is assumed that the zero suppression is turned on and that the data is 20 Bytes/channel hit. This assumption has been substantiated from a considerable volume of cosmic ray data. The VTX has 2064 channels in 10 crates, a full crate having 256 channels.

Our first problem was to determine the number of front end buffers needed. This was solved using an analytic model with 2 servers in series. The model is a single crate with multiple input buffers and a VBD with 2 buffers. The service time for the first server is the data transfer time across the backplane at 13 Mbytes/s. For the second server (VBD) it is the time to transfer the data over the data cable at 40 Mbytes/s. The event

size is taken from Table 1. Figure 2 shows the experiment dead time as a function of the event rate (the level 1 trigger rate) for 1, 2 and 3 buffers in the front end.

Layer	Mean	Sigma	Channels/crate	Crates
inner	4.5	2.1	128	2
middle	1.4	0.51	256	4
outer	1.6	0.62	208	4

Table I Expected VTX Hits / Channel for L1[2] Triggered Events

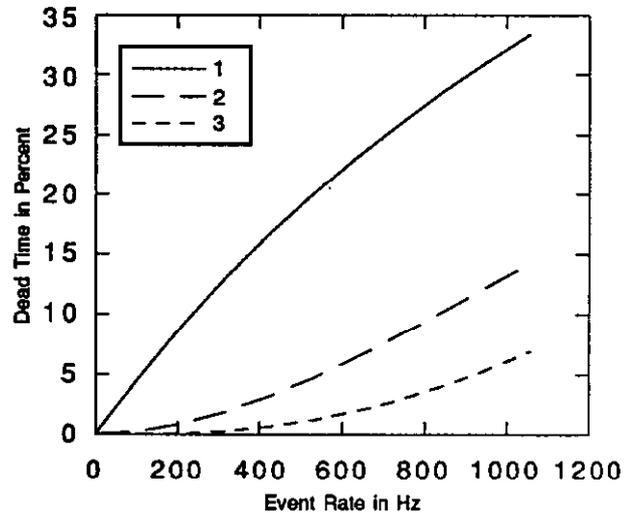


Figure 2. Experiment dead time as a function of event rate for 1, 2 and 3 front end buffers. The backplane rate is 13 Mbytes/s

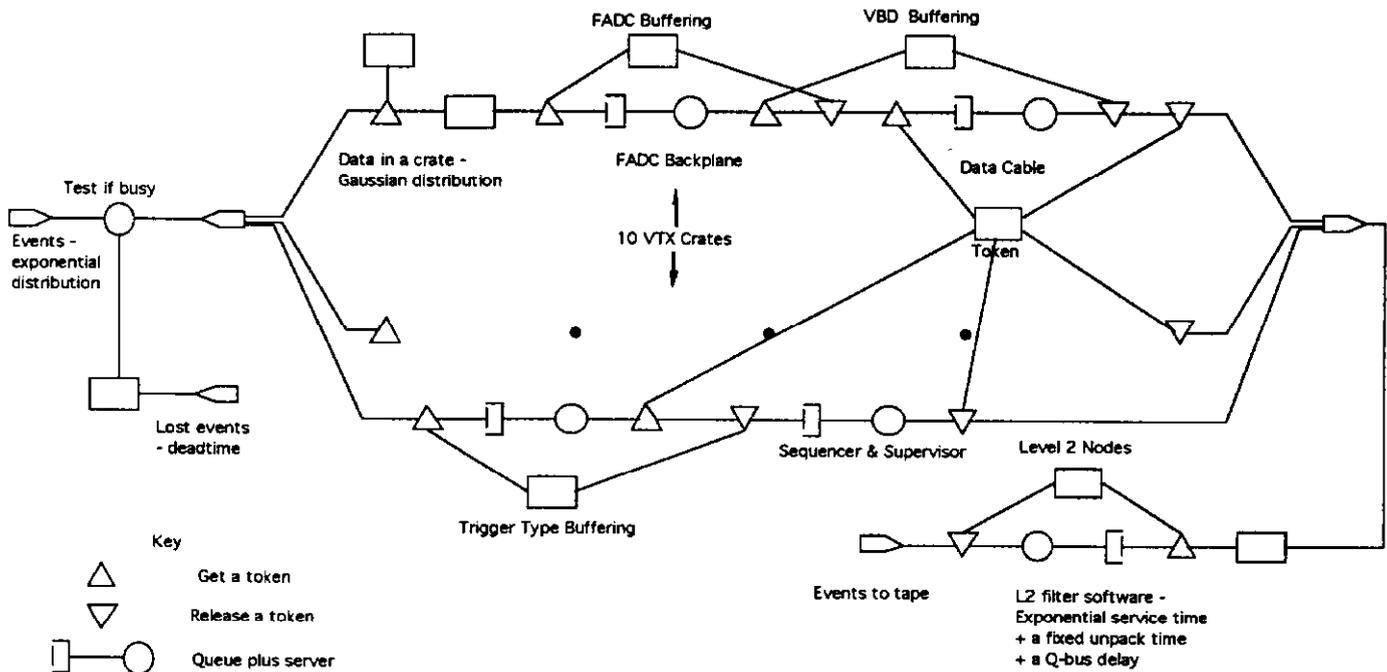
The design goal for DZero is a data rate between 200 and 400 Hz. As one can see from the figure, there is a large gain for going from 1 to 2 buffers but a rather small gain in going to 3. Therefore, we chose to build a double buffered system.

The next question asked was how fast should the backplane data transfer rate be. All crates transfer their data across the backplane in parallel but the readout over the data cable puts the crates in series. Thus, we need a model with multiple crates on a single data cable. This is beyond the scope of simple analytic models. We adopted the use of IBM's computer modeling package, Research Queueing Package Version 2 (RESQ)[3]. It is a Monte Carlo type of analysis tool that can handle arbitrarily complicated queuing models. It is only limited by computer time. We did not abandon the analytic models entirely. Whenever we made a substantial change to the computer model, we would run a simplified version that could be calculated by hand. This served as a running check on the correctness of the computer model. In one case where we could not compute an analytical solution, we wrote a small Monte Carlo program to check the computer model.

The queuing theory model of the VTX data cable is described in detail below and shown in Figure 3. It does not

include the entire data transfer system but only those elements that we believe contribute to significant delay.

Figure 3 RESQ Model for the Simulation of the VTX Readout



RESQ Model For The Simulation Of The VTX Readout

IV. A MONTE CARLO MODEL OF THE DATA ACQUISITION SYSTEM

A. RESQ Model

1. Front Ends

This is a single server queue with a queue length of one (double buffered). An analogy is a car wash with one space for washing a car and one parking spot out front. When the car wash is in use and the parking space occupied, customers must be turned away (the experiment has dead time). The arrival time of events is Poisson distributed. The service time (how long it takes to wash the car) is the time to transfer the data across the backplane. The event size which determines this time is given by the Gaussian distribution listed in Table I.

2. Data Transfer System.

The VBD is also double buffered. To carry the car wash analogy further, one can think of a wash station and wax

station as two separate but serial operations, each with one active station and two parking spaces in front. The data flow out of the VBD is the wax station. If the wax station is slow, all four spaces will fill up and customers will be turned away even though the car wash part may be done with its car. The event size determines the transfer time. Once the event is chosen for the front end, this time is also determined for the data cable since the same event is transferred across the backplane and through the VBD.

Several crates lie on each data cable. In the car wash analogy this corresponds to a large car wash with several lanes, each with a wash station (front end crates), but feeding into a single wax station (data cable). Each washing station has a parking spot in front of it and two afterwards. It should be remembered that in the case of the VTX readout there are 10 crates and all these crates are triggered at the same time, or in the analogy the cars arrive in groups of 10.

Figure 4 illustrates the analogy for 3 double buffered FADC crates and a single buffered VBD feeding a single data cable.

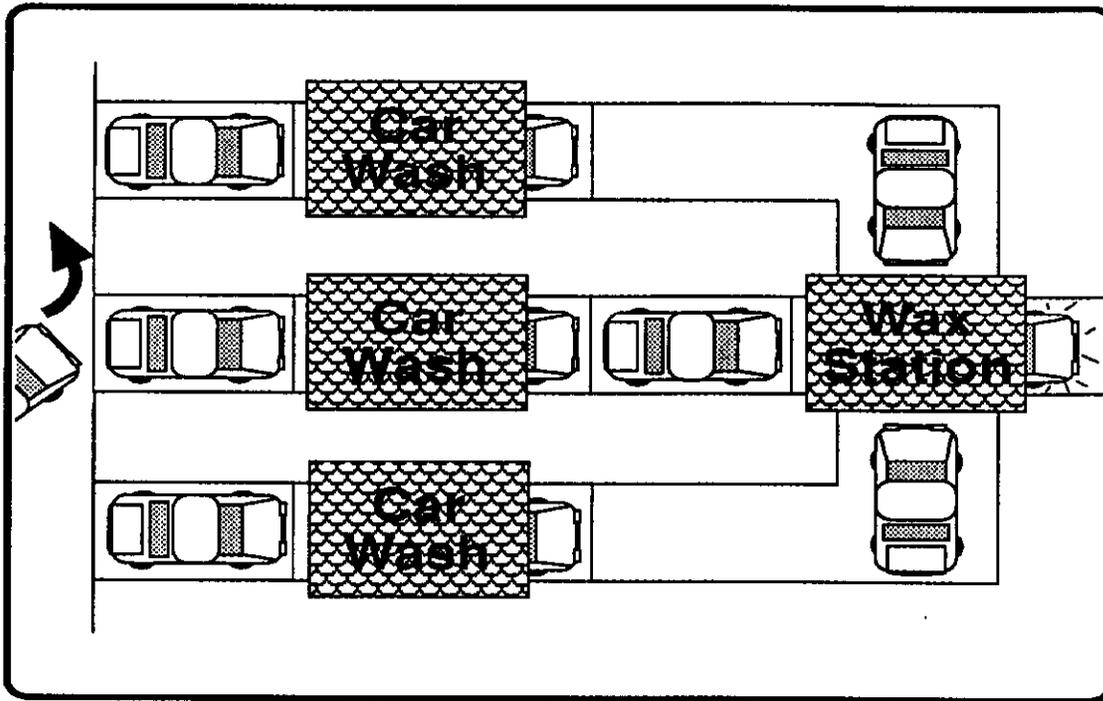


Figure 4 A Car Wash Analogy

E. Gonzalez 10/31/91

3. Level 2.

The level 2 nodes (50 MicroVAXs) correspond to a large number of servers. In the car wash analogy, they correspond to many cashiers at the end of the wax stations. Each cashier is very slow so many are required. The group of 10 cars that arrived together must all pay the same cashier. In our measurements the cashiers correspond to the Q-bus* transfer time from the local dual port memory to a MicroVAX. For actual running it will be the event processing time since new hardware will put the data directly into MicroVAX address space.

All level 2 nodes are on the data cable so some supervisor process must tell a node to accept an event. This is presently done in a supervisor MicroVAX. Also, not all crates are read out on every trigger so there must be something to formulate the token which controls the readout of the VBDs. This is done in a sequencer MicroVAX. There is only one supervisor for all the level 2 nodes and one sequencer for each data cable so there is no parallelism. However, there is a queue of trigger types that correspond to the data in the data cable queues. This parallels the buffering in the main data path. Without this queue, the data acquisition system is reduced to a 0 length queue.

The car wash analogy becomes a little contrived at this point but it is still useful. The manager walks to each of the cashiers in turn until a vacant one is found. He then walks to the wax station, prepares his materials, waxes the cars as they exit the wash and directs them to the chosen cashier. It is clear that these serial operations of finding a free cashier (analogous

to the supervisor finding a free MicroVAX) and the preparation of waxing materials (forming the token) is one of the bottlenecks in the current system. New hardware is now ready for installation that significantly speeds up this process.

4. Summary

In summary the model has 3 queues in series and one queue in parallel. Each crate is treated as one process, in other words, we ignore the individual channels. This is appropriate since typically all channels are processed in parallel, and the dominant time is backplane transfer time. We do only one data cable since all others are in parallel. We choose what we believe is the busiest data cable.

B. The Result of the RESQ Model

1. The Achieved Backplane Bandwidth

The design goal for the backplane transfer rate was 20 Mbytes/sec. For the Flash ADC system, a much simpler and less expensive design would limit the rate to 13 Mbytes/s. We wanted to know the effect on experiment dead time from using the lower transfer rate. Figure 5 shows the results of a simulation of the data flow from the 10 crates, across the backplanes to the VBDs and then down the data cable as controlled by the sequencer. The supervisor and sequencer delays have been ignored. The results show only a small difference between transfer rates of 13 and 20 Mbytes/s. At 200Hz from the L1 trigger and a backplane rate of 13 Mbytes/s the predicted dead time is 3% which is acceptable.

* Q-bus is a registered trade mark of Digital Equipment Corporation

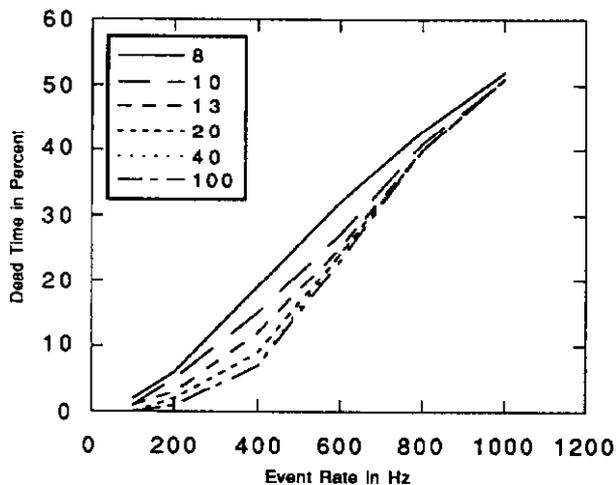


Figure 5. Data flow simulation for 10 VTX crates and a single data cable. Lines show different backplane data rates in Mbytes/s. The data is 20 bytes/channel hit with a 500ns overhead/channel. The data cable rate is 40 Mbytes/s. There is dual buffering in both the FADC and VBD.

2. The Process Time For The Online Software Trigger

The modeling study was extended to investigate the effect on dead time of the Software Trigger (L2). The time to make a trigger decision was assumed to be exponential in distribution. In the analogy this is the time taken by a cashier to accept payment from the 10 cars representing a triggered event.

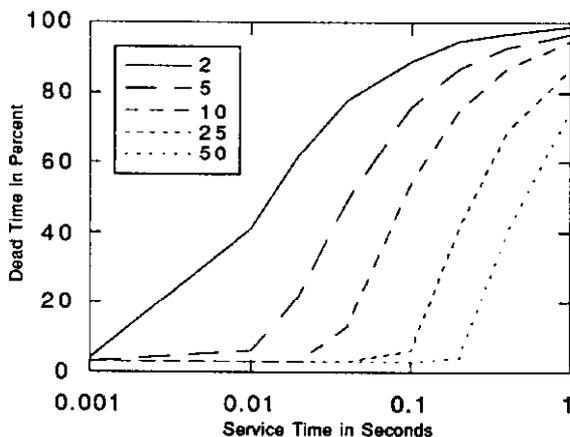


Figure 6. Data flow simulation for 10 VTX crates, a single data cable and a number of L2 nodes. The lines show different numbers of L2 nodes. The data is 20 bytes/channel hit with a 500ns overhead/channel. The FADC backplane data rate is 13 Mbytes/s and the data cable rate is 40 Mbytes/s. There is dual buffering in both the FADC and VBD. The event rate is 200 Hz.

Figure 6 shows the dead time for differing mean service times in L2 for a 200 Hz L1 trigger rate and the lines show different numbers of nodes.

3. The Sequencer and Supervisor Delays

A simulation was then run to investigate the constraint offered by the sequencer and supervisor delays and is shown in Figure 7. Again turning to the analogy, this is the delay for the manager to find an available cashier and to prepare for waxing. For these simulation runs the service time in L2 software was set to zero.

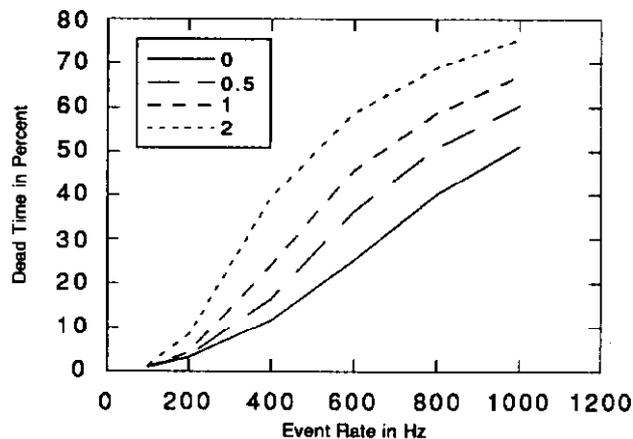


Figure 7. Data flow simulation for 10 VTX crates, a single data cable and L2 including the Supervisor delay. The lines show different sequencer + supervisor delays in ms. The data is 20 bytes/channel hit with a 500ns overhead/channel. The FADC backplane data rate is 13 Mbytes/s and the data cable rate is 40 Mbytes/s. There is dual buffering in both the FADC and VBD. The L2 service time is set to zero.

The simulation shows that this is a potential bottleneck in the DZero readout. Since this single queue is in parallel with the 10 VTX crates it is important that this delay is kept to a minimum. Significant improvements in this area are expected before the first physics run.

It is important that all L2 nodes handle all trigger types. Again turning to the analogy, the trigger types can be viewed as payment methods with only certain cashiers capable of dealing with credit cards or checks. The manager may well need to turn away a credit card customer as he knows that none of the card capable cashiers are free and would have to search longer to understand their availability.

V. A COMPARISON OF MODELING AND MEASURED PERFORMANCE

A. Performance Measurements of the DAQ System Used in the Cosmic Ray Test

Software was developed to allow downloading of either Monte Carlo events or selected bit patterns into the front-end data acquisition crates over the Token Ring control path. Once downloaded, the crates are triggered continuously to read out the same event. These events follow the same path as real data along the data cables through L2 and into the Host computer. By knowing exactly what was downloaded, we have an invaluable tool for testing long term integrity of the data acquisition system. We flag and dump any event read out that does not agree with what was downloaded; this enables us to pinpoint any questionable hardware in a timely fashion. We have also used this test procedure to measure the dead time of the currently installed system. The results of our tests are summarized below along with the predictions of our model.

We have downloaded data into 8 FADC crates (1536 channels) giving us an event size of 66,740 bytes, including the Trigger data block from L1. For these tests, we use a pulser trigger which we pre-scale to select the rate that we want. In our first test, we used 3 L2 nodes and trigger rates of 1-7 Hz before becoming bandwidth limited in passing data from L2 to the Host computer. Our second test involved only a single L2 node which was configured to pass a statistical sampling of the events read into L2 and on to the Host thus avoiding the above limitations. However, we did reach the bandwidth limitation of the Q-bus transfer from the dual port memory into the L2 node at a trigger rate of 9.5 Hz. Our third test used 2 L2 nodes without passing data to the HOST computer. These results are plotted in Figure 8 to compare with the simulations. Despite the model using a Poisson arrival distribution for the event rate, the agreement between measured system dead time and the predictions of the model is reasonable.

B. The DAQ System Currently Constrained by Q-bus

The DAQ system currently in use is constrained by the Q-bus transfer rate between the dual port memories at the end of each data cable and the L2 node processors. This will not be the case when the change is made to Multi Port Memories. To allow the simulation to be compared with measurements this delay was modelled and the results are shown in Figure 8. The transfer of data across Q-bus by the DZero software currently in use has been measured to be between 600 Kbytes/s and 1 Mbyte/s. A number of other changes were made to the simulation to match the actual test conditions.

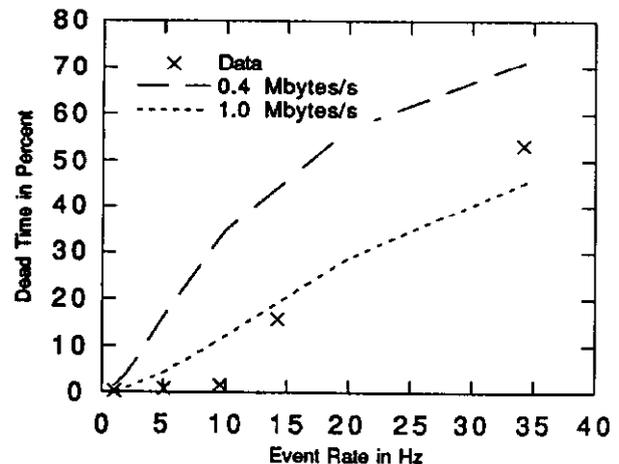


Figure 8. Data flow simulation for 8 VTX crates, a single data cable and L2 including both the Supervisor delay and the Q-bus constraint. The lines show the Q-bus transfer rate in Mbytes/s. The total data transferred is fixed at 67 Kbytes for the 8 crates plus the trigger. The FADC backplane data rate is 13 Mbytes/s and the data cable rate is 40 Mbytes/s. The Sequencer delay is 5.4 ms and the Supervisor delay is 2.5 ms. There are 2 L2 nodes with a zero service time.

VI. SUMMARY

Both the analytical and Monte Carlo studies of the DZero FADC crate readout has offered invaluable insights into the expected performance of the DAQ system. The application of queuing theory to the design of the FADC card and in particular to the choice of the level of buffering has led to a good design compromise between performance and price. We have also found that both models and analogies aid in understanding the design.

We started these models more than 5 years ago, and we continue to use them for system improvements and upgrades. We strongly recommend that a group starting out on a complex design develop models at the earliest stage.

VII. REFERENCES

- [1] D.Cutts, J.S.Hoftun, C.R.Johnson, R.T.Zeller "A Microprocessor Farm Architecture for High Speed Data Acquisition and Analysis", presented at the IEEE 1988 Nuclear Science Symposium, Orlando, Florida.
- [2] M.Abolins, D.Edmunds, P.Laurens, J.Linnemann, B.Pi "A High Luminosity Trigger Design for The Tevatron Collider Experiment D0" presented at the IEEE 1988 Nuclear Science Symposium, Orlando, Florida.
- [3] Charles Sauer, Edward MacNair and James Kurose. "The Research Queuing Package Version 2", RA138,RA139, IBM Research Division, San Jose, CA.